# Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow

**Julien Longhi**
Université de Cergy-Pontoise - CRTF
Julien.Longhi@u-cergy.fr

**Ciara R. Wigham**
Université Lumière Lyon 2 - ICAR
ciara.wigham@univ-lyon2.fr

The CoMeRe project (CoMeRe, 2014) aims to build a kernel corpus of computer-mediated communication (CMC) genres with interactions in the French language. Three key words characterize the project: variety, standards and openness. The project gathered mono- and multimodal, synchronous and asynchronous communication data from both Internet and telecommunication networks (text chat, tweets, SMSs, forums, blogs). A variety of interactions was sought: public or private interactions as well as interactions from informal, learning and professional situations.

Whereas some CMC data types were collected within the CoMeRe project, others had previously been collected and structured within different project partners' local research teams. This meant that the project had to overcome disparities in corpus compilation choices. For this reason, the CoMeRe project structured the corpora in a uniform way using the Text Encoding Initiative format (TEI, Burnard & Bauman, 2013) and decided to describe each corpus using Dublin Core and OLAC standards for metadata (DCMI, 2014; OLAC, 2008). The TEI model was extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse (Chanier *et al.*, 2014).

The term 'openness' also characterizes the project: The corpora have been released as open data on the French national platform of linguistic resources (ORTOLANG, 2013) in order to pave the way for scientific examination by partners not involved in the project as well as replicative and culumative research.

This poster presentation aims to give an overview of the corpus building process using, as a case study, a corpus of political tweets *cmr-polititweets* (Longhi *et al.*, 2014). The corpus stemmed from a local research project on lexicon (Digital Humanities and datajournalism, supported by the Fondation of Cergy-Pontoise University). It was built starting from seven French politicians from six different political parties. In order to generate political tweets, a set of lists citing these politicians was generated (7087 lists), and lists that have tweeted at least six times and for which the description contained the word 'politics' were selected (120 lists in total). Finally, 2934 tweets were recovered. In order to be sure that we selected politicians' tweets (and not, for example, those of journalists), only the accounts cited in more than 12 lists were considered; 205 politicians were tweeting. We took the last 200 tweets of each of the 205 accounts on 27 March 2014 (34,273 tweets). This allowed us to recover data that focused on the period between the two rounds of the 2014 municipal elections in France.

The poster will focus, firstly, on how features specific to Twitter were included and structured in the interaction space TEI model. We will exemplify how features including *hashtags* that label tweets so that other users can see tweets on the same topic, *at signs* that allow a user to mention or reply to other users and *retweets* that allow a user to repost a message from another Twitter user and share it with his own followers, were integrated into the model. Secondly, the poster will evoke some of the ethical and rights issues that had to be considered before publishing a corpus of tweets. Finally, the workflow & multi-stage quality control process adopted during the building of the corpus will be illustrated. This was an essential aspect considering that the corpus underwent format conversions: the local research team had initially structured the corpus in XML whilst the CoMeRe project applied the IS TEI model to the corpus.

The political tweets corpus is now structured and available online. Analyses have started to be carried out: some ideas have been launched in Djemili *et al.* (2014) but further analyses must adhere rigorously to methodologies stemming from the natural language processing (NLP) field.

## References

CoMeRe Repository (2014). Repository for the CoMeRe corpora [website], http://hdl.handle.net/11403/comere

Burnard, L. & Bauman, S. (2013). TEI P5: Guidelines for electronic text encoding and interchange. TEI consortium, *tei-c.org*. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C.R., Hriba, L., Longhi, J. & Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotation heterogeneous CMC genres, in Beißwenger, M., Oostdijk, N., Storrer, A & van den Heuvel, H. Building and Annotating Corpora of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics, *Journal of Language Technology and Computational Linguistics* (special issue). pp1-31.

http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

Djemili S., Longhi J., Marinica C., Kotzinos D. & Sarfati G.-E. (2014). « What does Twitter have to say about ideology », Konvens 2014 - *Workshop proceedings vol. 1* (NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media – Pre-conference workshop at Konvens2014) , Germany (2014), p.16-25.

DCMI (2014). Dublin Core Metadata Initiative. http://dublincore.org/

Longhi, J., Marinica, C., Borzic, B. & Alkhouli, A. (2014). Polititweets, corpus de tweets provenant de comptes politiques influents. In Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. http://hdl.handle.net/11403/comere/cmr-polititweets

OLAC. (2008). Best Practice Recommendations for Language Resource Description. *Open Language Archives Community*. University of Pennsylvania. http://www.languagearchives.org/REC/bpr.html

ORTOLANG (2013). Open Resources and TOols for LANGuage [website]. ATILF / CNRS - Université de Lorraine: Nancy, http://www.ortolang.fr